You Need to Get Yourself a PrivateGPT Right Now !!!!

Aditya Biradar

So "what is Privategpt ?" you might ask. PrivateGPT allows us to upload any document and then interact with that document will the power of LLM's (Large Language models) locally in your own device without the need of internet. It runs 100% locally so you can even upload your sensitive data without the fear of it getting stolen or leaked. You are free to chat with the LLM if you do not want to talk to your documents as it is open-source. It was originally created for linux and mac os but it is now possible to run it on windows taking into use the WSL (windows subsystem for linux).

Installation guide :

Prerequisites :

Install Anaconda from link: https://www.anaconda.com/download

Make sure to click the "add it to PATH environment variable " check box or add it manually when installation is done .

Step 1 : Creating a local environment and cloning repository.

Open the Anaconda Powershell Prompt in administrator mode. Use "cd 'path' " to change the directory to where you want to store the PrivateGPT and replace path with the path of the folder. Now type in the following commands to clone and set up environment for PrivateGPT:

• git clone https://github.com/imartinez/privateGPT

• cd privateGPT

• conda create -n privateGPT python=3.11

- conda activate privateGPT

Step 2 : Install Poetry.

Poetry simplifies the management of Python dependencies and packaging. Type in the following commands to install Poetry :

- conda install -c conda-forge pipx

- pipx install poetry

- set "PATH=%USERPROFILE%\.local\pipx\venvs\poetry\Scripts;%PATH%"

Now for installing the lamma LLM , UI and other dependencies using Poety:

- poetry install --extras "ui llms-ollama embeddings-ollama vector-stores-qdrant"

Step 3 : Running the Application.

Now we are ready to start the application, type in the following commands to get PrivateGPT started :

- cd scripts

- ren setup setup.py

- cd ..

- poetry run python scripts/setup.py

- set PGPT_PROFILES=local

- set PYTHONPATH=.

- poetry run python -m uvicorn private_gpt.main:app --reload --port 8001

It will take some time as it needs to download the LLM's Locally. After it finishes installing it will prompt " Application startup complete ". Now open your browser and enjoy your PrivateGPT at http://127.0.0.1:8001/ .

Now your PrivateGPT is running. You can futher accelerate the performance by GPU acceleration or using the chatgpt api. Since chatgpt api key is paid I will walk through the GPU Acceleration part

GPU Acceleration Guide :

Prerequisites :

Make sure you have a nvidia graphic card with cuda 118 support.

Install Visual Studio 2022.

Make sure the following components are selected:

☐ Universal Windows Platform development

☐ C++ CMake tools for Windows

Download the MinGW installer from https://sourceforge.net/projects/mingw/ .

☐ Run the installer and select the gcc component.

Step 1: Installing PyTorch with CUDA support:

• pip install torch==2.0.0+cu118 --index-url https://download.pytorch.org/whl/cu118

Step 2: Setting CMake arguments for llama-cpp-python:

• $env:CMAKE_ARGS='-DLLAMA_CUBLAS=on';

• poetry run pip install --force-reinstall --no-cache-dir llama-cpp-python

Step 3 : launch PrivateGPT with GPU support:

Reload and run the PrivateGPT with the following Commands :

• poetry run python -m uvicorn private_gpt.main:app --reload --port 8001

I have made my effort to compile the best and easiest working guide as of March 2024. I hope you will find this article usefull.